OCTOBER 15, 2016

ASSIGNMENT 1

PROJECT GROUP 11

SPARK USE CASE: 42

CMPE 272 – ENTERPRISE SOFTWARE PLATFORM

ADITI SHETTY
GitHub ID: shettyaditi

SURAJ KHURANA
GitHub ID: khurana3773

ARPITA DIXIT
GitHub ID: ArpitaDixit

NRUPESH PATEL
GitHub ID: Nrupesh29

Notebook Link: https://goo.gl/OaxMGQ

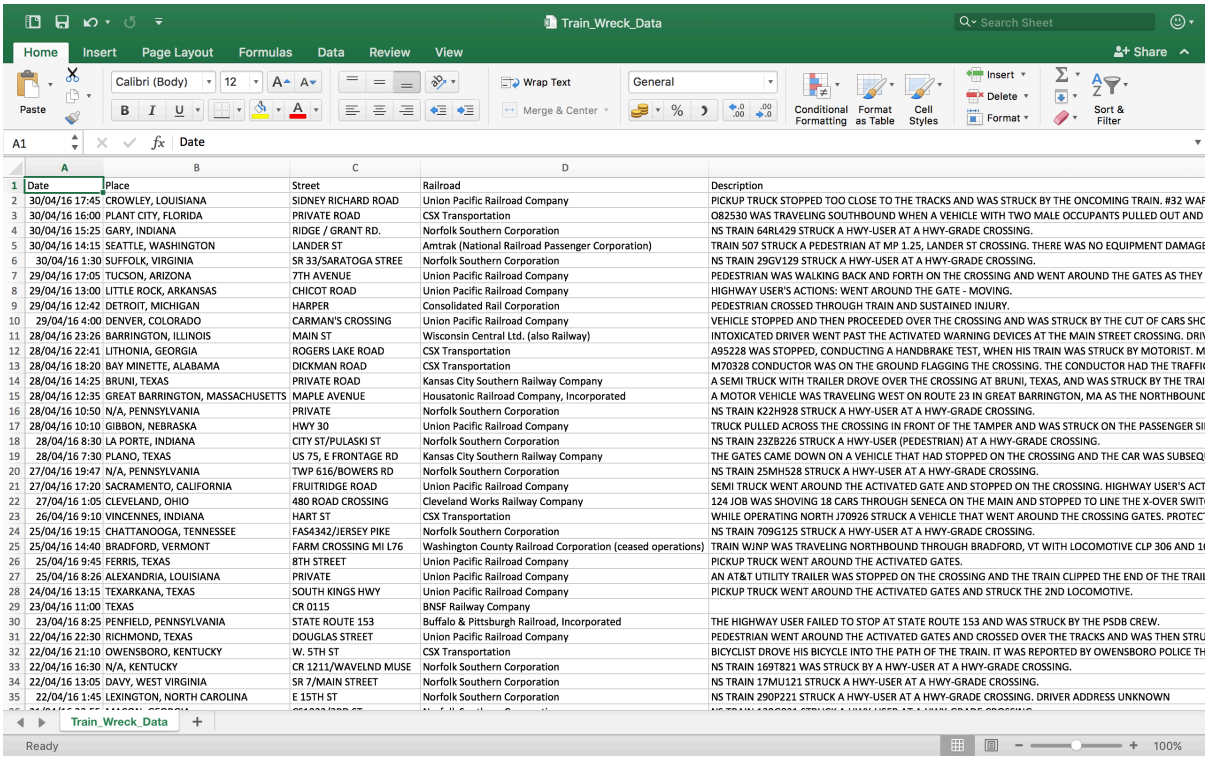
DESCRIPTION

(Use Case 42)

Use train wreck datasets http://www.trainwreckdb.com/ with spark to figure out what are the 10 most dangerous places for accidents.

DATA SOURCE

The following snapshot shows the data used from Train Wreck Dataset.



Train_Wreck_Data

Date	Place	Street	Railroad	Description
30/04/16 17:45	CROWLEY, LOUISIANA	SIDNEY RICHARD ROAD	Union Pacific Railroad Company	PICKUP TRUCK STOPPED TOO CLOSE TO THE TRACKS AND WAS STRUCK BY THE ONCOMING TRAIN. #32 WAS
30/04/16 16:00	PLANT CITY, FLORIDA	PRIVATE ROAD	CSX Transportation	O82530 WAS TRAVELING SOUTHBOUND WHEN A VEHICLE WITH TWO MALE OCCUPANTS PULLED OUT AND
30/04/16 15:25	GARY, INDIANA	RIDGE / GRANT RD.	Norfolk Southern Corporation	NS TRAIN 64RL429 STRUCK A HWY-USER AT A HWY-GRADE CROSSING.
30/04/16 14:15	SEATTLE, WASHINGTON	LANDER ST	Amtrak (National Railroad Passenger Corporation)	TRAIN 507 STRUCK A PEDESTRIAN AT MP 1.25, LANDER ST CROSSING. THERE WAS NO EQUIPMENT DAMAGE
30/04/16 1:30	SUFFOLK, VIRGINIA	SR 33/SARATOGA STREE	Norfolk Southern Corporation	NS TRAIN 29GV129 STRUCK A HWY-USER AT A HWY-GRADE CROSSING.
29/04/16 17:05	TUCSON, ARIZONA	7TH AVENUE	Union Pacific Railroad Company	PEDESTRIAN WAS WALKING BACK AND FORTH ON THE CROSSING AND WENT AROUND THE GATES AS THEY
29/04/16 13:00	LITTLE ROCK, ARKANSAS	CHICOT ROAD	Union Pacific Railroad Company	HIGHWAY USER'S ACTIONS: WENT AROUND THE GATE - MOVING.
29/04/16 12:42	DETROIT, MICHIGAN	HARPER	Consolidated Rail Corporation	PEDESTRIAN CROSSED THROUGH TRAIN AND SUSTAINED INJURY.
29/04/16 4:00	DENVER, COLORADO	CARMAN'S CROSSING	Union Pacific Railroad Company	VEHICLE STOPPED AND THEN PROCEEDED OVER THE CROSSING AND WAS STRUCK BY THE CUT OF CARS SHO
28/04/16 23:26	BARRINGTON, ILLINOIS	MAIN ST	Wisconsin Central Ltd. (also Railway)	INTOXICATED DRIVER WENT PAST THE ACTIVATED WARNING DEVICES AT THE MAIN STREET CROSSING. DRI
28/04/16 22:41	LITHONIA, GEORGIA	ROGERS LAKE ROAD	CSX Transportation	A95228 WAS STOPPED, CONDUCTING A HANDBRAKE TEST, WHEN HIS TRAIN WAS STRUCK BY MOTORIST. M
28/04/16 18:20	BAY MINETTE, ALABAMA	DICKMAN ROAD	CSX Transportation	M70328 CONDUCTOR WAS ON THE GROUND FLAGGING THE CROSSING. THE CONDUCTOR HAD THE TRAFFI
28/04/16 14:25	BRUNI, TEXAS	PRIVATE ROAD	Kansas City Southern Railway Company	A SEMI TRUCK WITH TRAILER DROVE OVER THE CROSSING AT BRUNI, TEXAS, AND WAS STRUCK BY THE TRAI
28/04/16 12:35	GREAT BARRINGTON, MASSACHUSETTS	MAPLE AVENUE	Housatonic Railroad Company, Incorporated	A MOTOR VEHICLE WAS TRAVELING WEST ON ROUTE 23 IN GREAT BARRINGTON, MA AS THE NORTHBOUND
28/04/16 10:50	N/A, PENNSYLVANIA	PRIVATE	Norfolk Southern Corporation	NS TRAIN K22H928 STRUCK A HWY-USER AT A HWY-GRADE CROSSING.
28/04/16 10:10	GIBBON, NEBRASKA	HWY 30	Union Pacific Railroad Company	TRUCK PULLED ACROSS THE CROSSING IN FRONT OF THE TAMPER AND WAS STRUCK ON THE PASSENGER SI
28/04/16 8:30	LA PORTE, INDIANA	CITY ST/PULASKI ST	Norfolk Southern Corporation	NS TRAIN 23ZB226 STRUCK A HWY-USER (PEDESTRIAN) AT A HWY-GRADE CROSSING.
28/04/16 7:30	PLANO, TEXAS	US 75, E FRONTAGE RD	Kansas City Southern Railway Company	THE GATES CAME DOWN ON A VEHICLE THAT HAD STOPPED ON THE CROSSING AND THE CAR WAS SUBSEQ
27/04/16 19:47	N/A, PENNSYLVANIA	TWP 616/BOWERS RD	Norfolk Southern Corporation	NS TRAIN 25MH528 STRUCK A HWY-USER AT A HWY-GRADE CROSSING.
27/04/16 17:20	SACRAMENTO, CALIFORNIA	FRUITRIDGE ROAD	Union Pacific Railroad Company	SEMI TRUCK WENT AROUND THE ACTIVATED GATE AND STOPPED ON THE CROSSING. HIGHWAY USER'S ACT
27/04/16 1:05	CLEVELAND, OHIO	480 ROAD CROSSING	Cleveland Works Railway Company	124 JOB WAS SHOVING 18 CARS THROUGH SENECA ON THE MAIN AND STOPPED TO LINE THE X-OVER SWIT
26/04/16 9:10	VINCENNES, INDIANA	HART ST	CSX Transportation	WHILE OPERATING NORTH J70926 STRUCK A VEHICLE THAT WENT AROUND THE CROSSING GATES. PROTEC
25/04/16 19:15	CHATTANOOGA, TENNESSEE	FAS4342/JERSEY PIKE	Norfolk Southern Corporation	NS TRAIN 709G125 STRUCK A HWY-USER AT A HWY-GRADE CROSSING.
25/04/16 14:40	BRADFORD, VERMONT	FARM CROSSING MI L76	Washington County Railroad Corporation (ceased operations)	TRAIN WJNP WAS TRAVELING NORTHBOUND THROUGH BRADFORD, VT WITH LOCOMOTIVE CLP 306 AND 1
25/04/16 9:45	FERRIS, TEXAS	8TH STREET	Union Pacific Railroad Company	PICKUP TRUCK WENT AROUND THE ACTIVATED GATES.
25/04/16 8:26	ALEXANDRIA, LOUISIANA	PRIVATE	Union Pacific Railroad Company	AN AT&T UTILITY TRAILER WAS STOPPED ON THE CROSSING AND THE TRAIN CLIPPED THE END OF THE TRAI
24/04/16 13:15	TEXARKANA, TEXAS	SOUTH KINGS HWY	Union Pacific Railroad Company	PICKUP TRUCK WENT AROUND THE ACTIVATED GATES AND STRUCK THE 2ND LOCOMOTIVE.
23/04/16 11:00	TEXAS	CR 0115	BNSF Railway Company	
23/04/16 8:25	PENFIELD, PENNSYLVANIA	STATE ROUTE 153	Buffalo & Pittsburgh Railroad, Incorporated	THE HIGHWAY USER FAILED TO STOP AT STATE ROUTE 153 AND WAS STRUCK BY THE PSDB CREW.
22/04/16 22:30	RICHMOND, TEXAS	DOUGLAS STREET	Union Pacific Railroad Company	PEDESTRIAN WENT AROUND THE ACTIVATED GATES AND CROSSED OVER THE TRACKS AND WAS THEN STRU
22/04/16 21:10	OWENSBORO, KENTUCKY	W. 5TH ST	CSX Transportation	BICYCLIST DROVE HIS BICYCLE INTO THE PATH OF THE TRAIN. IT WAS REPORTED BY OWENSBORO POLICE TH
22/04/16 16:30	N/A, KENTUCKY	CR 1211/WAVELND MUSE	Norfolk Southern Corporation	NS TRAIN 169T821 WAS STRUCK BY A HWY-USER AT A HWY-GRADE CROSSING.
22/04/16 13:05	DAVY, WEST VIRGINIA	SR 7/MAIN STREET	Norfolk Southern Corporation	NS TRAIN 17MU121 STRUCK A HWY-USER AT A HWY-GRADE CROSSING.
22/04/16 1:45	LEXINGTON, NORTH CAROLINA	E 15TH ST	Norfolk Southern Corporation	NS TRAIN 290P221 STRUCK A HWY-USER AT A HWY-GRADE CROSSING. DRIVER ADDRESS UNKNOWN

JUPYTER NOTEBOOK

Following snapshots show the Jupyter Notebook of our team for the selected spark use case.

In [1]:
```
def set_hadoop_config(credentials):
    prefix = "fs.swift.service." + credentials['name']
    hconf = sc._jsc.hadoopConfiguration()
    hconf.set(prefix + ".auth.url", credentials['auth_url']+'/v3/auth/tokens')
    hconf.set(prefix + ".auth.endpoint.prefix", "endpoints")
    hconf.set(prefix + ".tenant", credentials['project_id'])
    hconf.set(prefix + ".username", credentials['user_id'])
    hconf.set(prefix + ".password", credentials['password'])
    hconf.setInt(prefix + ".http.port", 8080)
    hconf.set(prefix + ".region", credentials['region'])
    hconf.setBoolean(prefix + ".public", True)
```

In [2]:
```
credentials = {
  'auth_url':'https://identity.open.softlayer.com',
  'project':'object_storage_afb98502_59e1_4eb1_a646_f3a3969f898d',
  'project_id':'0e0fc7828c5e4d2a92ec9aed289af4c6',
  'region':'dallas',
  'user_id':'6d3e88c5bf964f86b2e6bbb53d561d1f',
  'domain_id':'5c0e333aa3fd47d29a282818e646e8ed',
  'domain_name':'1141207',
  'username':'admin_b0a092623127b245412dec1a4bf9fe18b441700c',
  'password':"""kX..27t[pWEc}V[z""",
  'filename':'Train_Wreck_Data.csv',
  'container':'notebooks',
  'tenantId':'s68c-e03e15c2338850-5b7840cc6294'
}
```

In [3]:
```
credentials['name'] = 'keystone'
set_hadoop_config(credentials)
```

In [4]:
```
from __future__ import division
import numpy as np

from pyspark.sql import SQLContext
sqlContext = SQLContext(sc)

sc.addPyFile("https://raw.githubusercontent.com/seahboonsiew/pyspark-csv/master/pyspark_csv.py")
import pyspark_csv as pycsv

accidents = sc.textFile("swift://" + credentials['container'] + "." + credentials['name'] + "/Train_Wreck_Data.csv")

def skip_header(idx, iterator):
    if (idx == 0):
        next(iterator)
    return iterator

accidents_header = accidents.first()


accidents_header_list = accidents_header.split(",")
accidents_body = accidents.mapPartitionsWithIndex(skip_header)

accidents_body = accidents_body.filter(lambda line : len(line.split(","))>5)

accidents_df = pycsv.csvToDataFrame(sqlContext, accidents_body, sep=",", columns=accidents_header_list)
accidents_df.cache()
```

Out[4]:
```
DataFrame[Date: timestamp, Place: string, Street: string, Railroad: string, Description: string]
```

In [5]:
```
accidents_df.printSchema()
```

```
root
 |-- Date: timestamp (nullable = true)
 |-- Place: string (nullable = true)
 |-- Street: string (nullable = true)
 |-- Railroad: string (nullable = true)
 |-- Description: string (nullable = true)
```

In [6]:
```
accidents_df.take(1)
```

Out[6]:
```
[Row(Date=datetime.datetime(2016, 4, 30, 17, 45), Place=u'CROWLEY, LOUISIANA', Street=u'SIDNEY RICHARD ROAD', Railroad=u'Union Pacific Railroad Company', Description=u'PICKUP TRUCK STOPPED TOO CLOSE TO THE TRACKS AND WAS STRUCK BY THE ONCOMING TRAIN. #32 WARNING DEVICES: YIELD SIGN')]
```

In [7]:
```
accidents_df.count()
```

Out[7]:
```
541
```

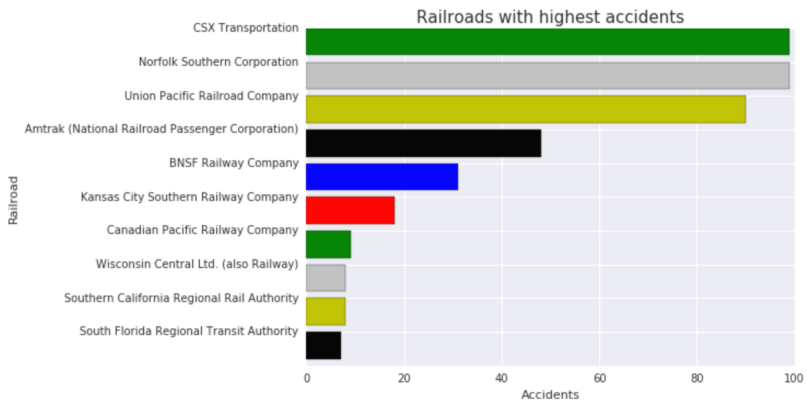
In [8]:
```
!pip install --user seaborn
```

```
Requirement already satisfied (use --upgrade to upgrade): seaborn in /gpfs/global_fs01/sym_shared/YPProdSpark/user/s68c-e03e15c2338850-5b7840cc6294/.local/lib/python2.7/site-packages
```

In [9]:
```
%matplotlib inline

import matplotlib.pyplot as plt

import matplotlib.patches as mpatches

import seaborn as sns
import pandas as pd
```

In [10]:
```
accidents_pd = accidents_df[['Street', 'Place', 'Railroad', 'Date', 'Description']].toPandas()

accidents_pd.columns = ['Street', 'Place', 'Railroad', 'Date', 'Description']
```

In [11]:
```
railroads = accidents_df.groupBy('Railroad').count().sort('count', ascending=False).toPandas()
colors = ['g','0.75','y','k','b','r']
railroads = railroads.ix[:9]
plt.barh(range(10),railroads['count'], color=colors)
plt.xlabel('Accidents')
plt.ylabel('Railroad')
plt.title('Railroads with highest accidents', size=15)
plt.yticks(range(10), railroads['Railroad'])
plt.gca().invert_yaxis()
plt.show()
```



In [12]:
```
place = accidents_df.groupBy('Place').count().sort('count', ascending=False).toPandas()
colors = ['g','0.75','y','k','b','r']
place = place.ix[:9]
plt.barh(range(10),place['count'], color=colors)
plt.xlabel('Accidents')
plt.ylabel('City, States')
plt.title('Places (City, States) with highest accidents', size=15)
plt.yticks(range(10), place['Place'])
plt.gca().invert_yaxis()
plt.show()
```